Two Competing Models of How People Learn in Games

Ed Hopkins*
Department of Economics
University of Edinburgh
Edinburgh EH8 9JY, UK

December, 2000

Abstract

Reinforcement learning and stochastic fictitious play are apparent rivals as models of human learning. They embody quite different assumptions about the processing of information and optimisation. This paper compares their properties and finds that they are far more similar than were thought. In particular, the expected motion of stochastic fictitious play and reinforcement learning with experimentation can both be written as a perturbed form of the evolutionary replicator dynamics. Therefore they will in many cases have the same asymptotic behaviour. In particular, they have identical local stability properties at mixed equilibria. The main identifiable difference between the two models is speed: stochastic fictitious play gives rise to faster learning.

Journal of Economic Literature classification numbers: C72, D83.

Keywords: Games, Reinforcement Learning, Fictitious Play.

*This paper arose out of extensive discussions with Tilman Börgers. It was completed while visiting the Economics Department at the University of Pittsburgh, which I thank for its hospitality. I have benefitted from comments from Josef Hofbauer, John Duffy, Martin Posch, Al Roth, Tatiana Kornienko, Peyton Young, three referees, and the Editor, Drew Fudenberg. Errors remain my own. E.Hopkins@ed.ac.uk, http://www.ed.ac.uk/~ehk

1 Introduction

What is the best way to model how people learn to play games? This topic has been studied intensively in recent years from an empirical standpoint. There has been much progress in fitting the predictions of learning models to data on individual behaviour in experiments. But there has been some disagreement about what obtains the best fit: reinforcement learning, stochastic fictitious play or a model which encompasses or combines the two. See for example, Erev and Roth (1998), Camerer and Ho (1999), Sarin and Vahid (1998), Feltovich (2000), Salmon (1999), Blume et al. (2000).

This paper takes a different approach. While remaining agnostic as to which model or models best describe actual human learning behaviour, the theoretical properties of reinforcement learning and stochastic fictitious play are compared. The models differ on two levels, what information agents use and whether agents optimise given that information. Nonetheless, is possible to take the result of Camerer and Ho (1999) a stage further: there is more than just a family resemblance between the two models. As will be seen, both models can be considered as noisy versions of the evolutionary replicator dynamics. This has the result that by the choice of an appropriate noise function, one can construct a pure reinforcement model, with no optimisation and which throws information away, which has expected motion identical (to a positive scalar transformation) to that of the most commonly used form of stochastic fictitious play.

This means that, first, the equilibria of a perturbed reinforcement learning model can be identical to those of stochastic fictitious play. That is, if the two models do converge they will converge to the same point. A second important result is on local stability of these learning schemes. If any mixed equilibrium is locally stable (unstable) for all forms of stochastic fictitious play, it is locally stable (unstable) for perturbed reinforcement learning. A general principle might be that if a game has sufficient structure that we can be sure of the asymptotic behaviour of fictitious play, then reinforcement learning will exhibit the same behaviour.

This paper is also able to resolve some apparently contradictory claims. For example, Erev and Roth (1998) argue that the experimental data on play in 2×2 games with a unique mixed strategy equilibrium does not seem to support traditional equilibrium analysis. Simply put, the experimental subjects often did not play the unique Nash equilibrium even in these simple games. In the short run, the dynamics may seem to move away from equilibrium. Even after a significant length of time, play is not near Nash equilibrium. Indeed, it has been known for some time that the expected motion of reinforcement learning is given by the evolutionary replicator dynamics, see for example, Börgers and Sarin (1997), Posch (1997). The replicator dynamics do not converge to equilibrium in this class of games, a result that Posch (1997) uses to show that the basic model of reinforcement learning typically will not do so either.

In contrast, Fudenberg and Kreps (1993), Benaïm and Hirsch (1999) have shown that stochastic fictitious play converges in these games. However, convergence is not to the Nash equilibrium. Theorists traditionally think of the possibility of agents making mistakes, and the noise this generates, as vanishingly small. In contrast, the literature on quantal response equilibria following from McKelvey and Palfrey (1995) has emphasised that the equilibria of a game with perturbed payoffs are qualitatively different from the original Nash equilibria. Similarly, once noise is added to learning, whether reinforcement learning or fictitious play, the steady states of the learning process will not be identical to Nash equilibria. In the particular case of 2×2 games with a unique Nash equilibrium, reinforcement learning with experimentation converges to a perturbed equilibrium in a similar manner to stochastic fictitious play. But since the point to which learning converges is not the Nash equilibrium, these convergence results are not in conflict with the experimental data.

Reinforcement or stimulus-response learning is a very simple application of the principle that actions that have led to good outcomes in the past are more likely to be repeated in the future. Agents have a probability distribution over possible actions. When an action is chosen, the probability of that action being taken again rises in proportion to the realised payoff. The action has been "reinforced". Note the very low level of information or processing ability necessary to implement such an algorithm. In the context of game-playing, an agent does not need to know the structure of the game, to calculate best responses or even to know that a game is being played. In contrast, fictitious play assumes agents optimise given their beliefs about the play of their opponents, even if those beliefs are formed adaptively.

This emphasises that there are in fact two fundamental issues which divide the differing models considered here. First, information: are agents sophisticated enough to work out what payoffs they might have received if they had played some strategy other than they actually did? This is what Camerer and Ho (1999) call "hypothetical reinforcement" and it is present in classical fictitious play. In its absence there will be what Erev and Roth (1998) call "force of habit". That is, agents will be biased to repeat actions they have taken in the past. Second, optimisation: do agents maximise given their beliefs as in fictitious play, or do they use a more probabilistic rule as in reinforcement learning?

The question is whether these differences in the models lead to qualitatively different predictions on the form that learning should take. As noted the expected motion of stochastic fictitious play differs from that of reinforcement learning by a positive factor. This factor is in fact given by the parameter that determines the level of optimisation in stochastic fictitious play. Thus, differences in the level of optimisation can have a direct effect on the speed of convergence. In contrast, and rather surprisingly, the analysis finds no similar direct role for the additional information used in fictitious play.

This paper is structured in the following way. Section 2 introduces and compares

reinforcement learning and fictitious play. Section 3 investigates the dynamics of the two models and shows how the expected motion of stochastic fictitious play can be expressed as a form of replicator dynamic. Section 4 analyses the impact of noise on reinforcement learning. Section 5 outlines the basics of stochastic approximation theory and its application to these two models of learning. Section 6 compares the perturbed equilibria of the two models and gives some local and global stability results. Section 7 concludes.

2 Two Competing Models of Learning

This paper examines learning in the context of 2-person normal form games. This section introduces two rival models, reinforcement learning and fictitious play. There are two agents A and B who play a game repeatedly at discrete time intervals, indexed by n. The first player, A, has N strategies, B has M. In period n, A's mixed strategy is written $x_n \in S_N$, and the strategy of B, $y_n \in S_M$ where S_N is the simplex $\{x = (x_1, ..., x_N) \in I\!R^N : \Sigma x_i = 1, x_i \geq 0, \text{ for } i = 1, ..., N\}$. Let A be the $N \times M$ payoff matrix for the first player, with typical element a_{ij}, and B be the $M \times N$ payoff matrix for the second player with typical element b_{ji}. Expected payoffs for A will be $x \cdot Ay$, and for B, $y \cdot Bx$.

As Erev and Roth (1998) and Camerer and Ho (1999) observe, it is possible to compare fictitious play and reinforcement learning directly by working in terms of "attractions" or "propensities". Here it is assumed that if A has N strategies, then she has N propensities, which in period n will be given by a vector $q_n^A = (q_{1n}^A, ..., q_{Nn}^A)$. Then the state of the system can be written $q_n = (q_n^A, q_n^B)$. Fictitious play and reinforcement learning differ both in terms of how these propensities determine choice probabilities and how these propensities are updated as a result of realised play.

First, consider what Roth and Erev (1995), Erev and Roth (1998) call the basic reinforcement learning model. The easiest way of describing how the algorithm works is to imagine that an agent k in period n has an urn containing a total of Q_n^k balls of N different colours with q_{in}^k balls of the ith colour. Each period the agent draws one ball at random (with replacement) from the urn and takes the action corresponding to the colour drawn. The probability of the player A, respectively B, taking his ith action in period n is therefore,

$$x_{in} = \frac{q_{in}^A}{\sum_{j=1}^N q_{jn}^A} = \frac{q_{in}^A}{Q_n^A}, \; y_{in} = \frac{q_{in}^B}{\sum_{j=1}^M q_{jn}^B} = \frac{q_{in}^B}{Q_n^B}. \tag{1}$$

Strategies with higher propensities are played with higher probability. It is always assumed that all initial propensities are strictly positive, so at all times, there will be a positive probability of a strategy being picked.

To specify a learning model, one also needs an updating rule. For the basic reinforcement model, one can imagine after each period of play that the agent adds to the urn a number of balls equal to the payoff received of the colour corresponding to the action taken. If player A takes action i, and player B chooses j in period n, describe this as the event ij. The function f_{ij} is known as the event operator associated with the event ij. The propensities are updated thus,

$$\begin{array}{ll} q^A_{in+1} & = f_{ij}(q^A_{in}) = q^A_{in} + a_{ij} \\ q^A_{kn+1} & = f_{ij}(q^A_{kn}) = q^A_{kn} \qquad \text{for all } k \neq i. \end{array} \tag{2}$$

That is, only the ith propensity is changed. The justification is that since the actions other than i were not chosen, the payoff they would have earned is not observed.

In contrast the choice rule for fictitious play, for player one for example, is given by

$$x_{in} = 1 \text{ if } q^A_{in} = \max q^A_n, \text{ else } x_{in} = 0. \tag{3}$$

That is, the strategy that appears the best is played with probability one.[1] The updating rule for fictitious play is also different and is based on the following argument. If one's own payoff matrix is known and the action of one's opponent is observable, a player could calculate what he would have received had he chosen some other action. In particular, if player A were to observe that B is playing her jth strategy, under fictitious play all propensities would be updated thus,

$$q^A_{in+1} = f_{ij}(q^A_{in}) = q^A_{in} + a_{ij} \text{ for } i = 1, ..., N. \tag{4}$$

That is, A reasons that if he had chosen action k, given B's choice of j, he would have received a_{kj} and he updates his kth propensity appropriately. This has been called "hypothetical reinforcement".

Fictitious play is often presented slightly differently with players holding beliefs about opponents' play based on the frequency of opponents' past choices. Let $u_n \in S_N$ be the vector of relative frequencies of the actions of the first player up to period n. That is, if after 100 rounds of play A has played the first of two strategies 30 times, then $u_n = (0.3, 0.7)$. Let $v_n \in S_M$ be the vector of the relative frequencies of the choices of the second player. Then, if player A chooses her strategy x_n to maximise $x_n \cdot Av_n$, this will be asymptotically equivalent to the choice rule (3). To see this, note that given (4), $q^A_{in} = q^A_{i0} + n(Av_n)_i$.[2] That is, both q^A_{in}/n and $(Av_n)_i$ give an estimated return to strategy i based on the opponent's past play.

Note that in fictitious play whether in terms of beliefs or propensities, updating by A is independent of the choices made by A. In contrast, the reinforcement learning rule

[1] Implicit here are additional assumptions to ensure that there are no ties for first place. I don't detail them as ties are not an issue in stochastic fictitious play which is the main object of interest and which is described later on in this section.

[2] However, this relationship does not hold if the reinforcement learning updating rule (2) is used. This emphasises that it is difficult to express reinforcement learning in terms of beliefs.

(2) exhibits what Erev and Roth call "force of habit". The actions which are chosen more frequently are reinforced more frequently. Force of habit is important in what it indicates about agents' processing of information. When information is available about choices of opponents, the standard reinforcement updating rule (2) throws this information away. The evidence from experiments as to what people actually do is mixed. Erev and Roth's (1998) detection of force of habit in the learning behaviour exhibited in their data is matched by Van Huyck et al. (1997) who find that force of habit is statistically insignificant in data from their experiments. Camerer and Ho (1999) claim the data supports an intermediate case. Furthermore, the differences between the two learning models in their treatment of information is in practice blurred. In fact, both updating rules have been used in the reinforcement learning literature (see Vriend, 1997) and there has been more than one attempt to specify a fictitious play-like learning process for use when opponents' actions are unobservable (see Fudenberg and Levine, 1998, Ch4; Sarin and Vahid, 1998).

There is one other potential difference between the two types of learning. In the basic reinforcement model if realised payoffs are negative, then this may lead to one of the propensities becoming negative and the probabilities x, y will no longer be defined. To cure this technical problem it is necessary (and in this paper, it is assumed) that all payoffs are strictly positive. Responses to this problem vary. Rustichini (1999) simply states that it is without any loss of generality to assume all payoffs positive. However, Börgers and Sarin (1997) argue that in this context payoffs should not be interpreted as von Neumann-Morgenstern utilities, rather they should instead be considered as "parameterizations of players' responses to experiences". Erev and Roth (1998) take a slightly different approach. The actual payoffs players receive may be positive or negative but it is assumed that players update their propensities with reinforcements which are always positive. Reinforcements differ from payoffs only by the addition of a constant.

The problem with fictitious play in contrast is that it is entirely deterministic. On a theoretical level, it means that convergence to mixed strategy equilibria is problematic. On an empirical level, it seems to obtain a very poor fit to, for example, the behaviour of experimental subjects. As a result of both factors, stochastic or smooth fictitious play has been increasingly popular. This is where the standard fictitious play updating rule is used but the deterministic choice of a best response is replaced by a stochastic choice rule. Remember that the fictitious play rule picks out the strategy with the highest propensity, which is equivalent to choosing the strategy with the highest historical payoff or choosing x to maximise $x \cdot Av_n$, where Av_n is the vector which describes A 's historical payoffs given B's past choices v_n. As set out in Fudenberg and Levine (1998, Chapter 4), it is possible to reconsider the maximisation problem when the player's payoffs are subject to noise, so that instead A chooses x to maximise

$$x \cdot Av_n + \lambda\phi(x),$$

where λ is a scaling factor for the perturbation $\phi(x)$.

(i) $\phi(x)$ is strictly concave and ϕ'', the matrix of second derivatives of ϕ with respect to x, is negative definite.

(ii) The gradient of ϕ becomes arbitrarily large near the boundary of the simplex, i.e., $\lim_{x\to\partial S_N} |\phi'(x)| = \infty$.

Then it is certain there exists a unique solution to the following first order conditions for a maximum,

$$x = (\phi')^{-1}\left(-\frac{1}{\lambda}Av_n\right) = \overline{BR}(v_n). \tag{5}$$

$\overline{BR}$ is thus a perturbed best response function, with λ as a noise parameter. As it drops to zero, the above rule approaches the standard fictitious play rule (3) and will pick out the strategy with the highest expected return with probability one. However, high values of λ, that is, lots of noise, will mean the probability of a best response will be much decreased.

The most commonly used procedure is to set $\phi(x) = -\sum x_i \log x_i$ and consequently the deterministic choice of strategy for player A is replaced by,

$$x_{in} = \frac{\exp \beta(Av_n)_i}{\sum_{j=1}^N \exp \beta(Av_n)_j} = \overline{BR}_i^e(v_n), \tag{6}$$

where the "e" superscript is for exponential, and I have written $\beta = 1/\lambda$ for concision. The other particular functional form that has been popular in the literature is

$$x_{in} = \frac{(Av_n)_i^\beta}{\sum_{j=1}^N (Av_n)_j^\beta} = \overline{BR}_i^p(v_n), \tag{7}$$

where the "p" superscript is for power. If $\beta = 1$, then this rule is similar to the reinforcement learning choice rule. It is possible to fit this form to experimental data and use estimates of β to test between reinforcement learning and stochastic fictitious play.[3]

3 Dynamics

Each of the two models has been introduced in terms of a choice rule and an updating rule. The two rules together define a discrete time stochastic process. In investigating these dynamics, it turns out that continuous time deterministic dynamics will be a useful tool. The link, which will become apparent in Section 5, is the theory of stochastic approximation. In particular, it will be useful to employ as a benchmark

[3]There is the problem, however, that this rule could not arise as the result of the maximisation of a perturbed payoff function as considered here. See Hofbauer and Hopkins (2000).

the evolutionary replicator dynamics. These dynamics have been analysed extensively (see for example, Hofbauer and Sigmund, 1998). It is possible to write them in vector form[4] as

$$\dot{x} = R(x)Ay, \dot{y} = R(y)Bx, \tag{8}$$

where again $x \in S^N$, $y \in S^M$ and $R(\cdot)$ is a symmetric positive semi-definite matrix which we can refer to as the replicator operator. The ith diagonal element of $R(x)$ is $x_i(1 - x_i)$ and the jth element of the ith row is $-x_i x_j$. The link between these dynamics and reinforcement learning is well known and has been explored in Börgers and Sarin (1997), Posch (1997) and Rustichini (1999).

3.1 Reinforcement Learning

The basic reinforcement learning model is defined by the choice rule (1) and the updating rule (2). However, the real interest is in the evolution of (x_n, y_n), that is, the players' mixed strategies. It turns out (calculations in the Appendix) that the expected change in (x_n, y_n) can be written as

$$E[x_{n+1}|q_n] - x_n = \frac{R(x_n)Ay_n}{Q_n^A} + O(\frac{1}{(Q_n^A)^2}),\ E[y_{n+1}|q_n] - y_n = \frac{R(y_n)Bx_n}{Q_n^B} + O(\frac{1}{(Q_n^B)^2}), \tag{9}$$

where $R(\cdot)$ is the replicator operator that we have just introduced. Note first that the rate of change of x_i is decreasing in the sum of the propensities, Q_n^A, and that Q_n^A increases each period by a stochastic increment equal to the realised payoff. As this is strictly positive by assumption, both Q_n^A and Q_n^B are strictly increasing with order n. We refer to $1/Q_n^k$ as the *step size* of player k's learning process. Second, the expected motion of x is approximately equal to the evolutionary replicator dynamics (8).

It is worth remarking that there exist other forms of reinforcement learning. For example, Börgers and Sarin (1997) employ the Cross model. Both models have an expected motion close to that of the replicator dynamics, but they differ in terms of their step size. Here, as noted, the there is no unique step size to the learning process, rather each player having her own, $1/Q_n^A$ and $1/Q_n^B$ respectively, which are both decreasing over time. In the Cross model the step size is exogenous and fixed at a constant level, say γ. Börgers and Sarin (1997) show that in the limit as γ is exogenously reduced to zero the Cross learning process approaches the replicator dynamics (8). Here, our dual step sizes reduce to zero endogenously as time elapses.

[4]The replicator dynamics are more commonly written out element by element with, for example, $\dot{x}_i = x_i((Ay)_i - x \cdot Ay)$ for $i = 1, ..., N$ replacing $\dot{x} = Q(x)Ay$. But the vector form is more convenient in this context as will become apparent.

3.2 Stochastic Fictitious Play

The choice rule given by the perturbed best response function (5) combined with the updating rule (4) define a stochastic learning process. Rather than looking at the evolution of the choice probabilities (x_n, y_n), it is more usual to work with the historical frequencies of choices. More specifically, as for example, Benaïm and Hirsch (1999) calculate,

$$E[u_{n+1}|u_n, v_n] - u_n = \frac{\overline{BR}(v_n) - u_n}{n+1}, \; E[v_{n+1}|u_n, v_n] - v_n = \frac{\overline{BR}(u_n) - v_n}{n+1}. \tag{10}$$

The step size here is exactly $1/(n+1)$ which reflects the fact that (u_n, v_n) are time averages.[5]

In order to compare stochastic fictitious play and reinforcement learning directly, it is convenient to adopt the novel approach of looking at current mixed strategies rather than the historical frequencies used in (10) above. For example, it is possible to obtain the expected change in the probability A places on her ith strategy by summing over the j possible actions of her opponent (remember updating in fictitious play is independent of one's own choice),

$$E[x_{in+1}|x_n, y_n] - x_{in} = \sum_{j=1}^{M} y_{jn} \left(\overline{BR}_i \left(f_{ij}(Av_n) \right) - \overline{BR}_i(Av_n) \right). \tag{11}$$

Note that from (10) above the step size of the change in v is of order $1/n$. Hence, it is possible to apply the following approximation:

$$E[x_{in+1}|x_n, y_n] - x_{in} = \sum_{j=1}^{M} y_{jn} \left(\frac{\partial \overline{BR}_i}{\partial Av_n} \cdot \left(f_{ij}(Av_n) - Av_n \right) \right) + O(\frac{1}{n^2}), \tag{12}$$

which in turn leads to the following result.

Proposition 1 *Stochastic fictitious play with updating rule (4) and choice rule (5) defines a stochastic process for the choice probabilities (x_n, y_n) of the two players given by*

$$\begin{aligned} x_{n+1} - x_n &= \tfrac{\beta}{n+1} P(x_n)(Ay_n + \lambda \phi'(x_n)) + \tfrac{\eta_n^A}{n+1} + O(\tfrac{1}{n^2}) \\ y_{n+1} - y_n &= \tfrac{\beta}{n+1} P(y_n)(Bx_n + \lambda \phi'(y_n)) + \tfrac{\eta_n^B}{n+1} + O(\tfrac{1}{n^2}), \end{aligned} \tag{13}$$

where $\beta = 1/\lambda$, $P(\cdot) = \overline{BR}'(\cdot) = -(\phi'')^{-1}(\cdot)$ is a symmetric positive semi-definite matrix function and η_n^A, η_n^B are random variables having expectation zero conditional on (x_n, y_n).

[5] If the original fictitious play choice rule (3) were employed instead, $\overline{BR}(\cdot)$ would be replaced in (10) by $BR(\cdot)$, the original best response correspondence. See, for example, Fudenberg and Levine (1998, Ch2).

Proof: In the Appendix. ∎

What is the advantage of this approach? First, it is worth remarking that it makes little difference whether one analyses stochastic fictitious play in terms of marginal or historical frequencies. Fudenberg and Kreps (1993) and Benaïm and Hirsch (1999) have already shown that convergence of the process in historical frequencies to an equilibrium point implies convergence of the marginal frequencies (x, y). But equally if the process in current choice frequencies converges to any point, then, by the law of large numbers, the time average of play must converge to that point also.

Second, this result shows that expected motion of the stochastic process (13) is actually a noisy version of the replicator dynamics produced by the "dumb" reinforcement model. To see this, note that it is possible to generalise the replicator dynamics (8) by replacing the replicator operator $R(\cdot)$ by an arbitrary symmetric positive semi-definite matrix function $P(\cdot)$ but while retaining many of their properties (Hopkins, 1999a, b). In fact, in the particular case of the exponential choice version of fictitious play (6) so that for example $\partial \overline{BR}_i^e / \partial (Av_n)_i = \beta x_i (1 - x_i)$, then $P(\cdot) = R(\cdot)$.[6] That is, for λ close to zero, the expected motion of stochastic fictitious play in current choice probabilities (13) given the exponential choice rule will be close to a positive multiple, β, of that of reinforcement learning.

More generally the fact that deterministic part of (13) is multiplied by the factor $\beta = 1/\lambda$ implies that the closer to optimisation agents are, the faster they learn. The other major difference between (13) above and the equivalent expression for reinforcement learning (9) is the additional term $\lambda \phi'(\cdot)$ which is best interpreted as a noise term. First, this is simply because its magnitude is directly proportional to the parameter λ. Second, because of the assumption (ii) on ϕ, it prevents any choice probability falling to zero. If the expected motion of reinforcement learning and stochastic fictitious play differ only in terms of noise, it suggests that a suitable perturbed version of reinforcement learning would generate the same expected motion as that of stochastic fictitious play. It will be seen that this is the case.

4 Perturbed Reinforcement Learning

We have seen how in fact that with the introduction of noise, the expected motion of fictitious play becomes a form of noisy replicator dynamic. The introduction of noise to reinforcement learning, as we will now see, has a similar result. Erev and Roth (1998) introduce what they call experimentation to the basic reinforcement learning model by assuming there is some reinforcement for all propensities, not just for the one corresponding to the action taken. That is, when player A takes action i, and B

[6] A relationship between the exponential choice rule and the replicator dynamics is also found in Rustichini (1999).

takes j, for some small $\lambda > 0$, updating rule (2) is replaced by

$$\begin{array}{ll} q^A_{in+1} = q^A_{in} + (1-\lambda)a_{ij} & \\ q^A_{kn+1} = q^A_{kn} + \frac{1}{N-1}\lambda a_{ij} & \text{for all } k \neq i. \end{array} \tag{14}$$

This specification might capture the idea of experimentation in that even though a strategy is currently performing poorly, it still receives some reinforcement and will not be entirely forgotten. In any case, if all payoffs are positive, this noise and/or experimentation will prevent the probability of taking any action, even if dominated, falling to zero.

Given that our knowledge as to which learning model best describes human behaviour is limited, the question as to what form noise should take is even more murky. As the specification chosen by Erev and Roth depends on the payoff earned, its expected motion will be function of both x and y. This makes it difficult to analyse for games larger than 2×2. The following simpler alternative to (14) is rather more tractable and no less plausible:

$$\begin{array}{ll} q^A_{in+1} = q^A_{in} + a_{ij} + \lambda & \\ q^A_{kn+1} = q^A_{kn} + \lambda & \text{for all } k \neq i. \end{array} \tag{15}$$

That is, all propensities are reinforced by a small amount.

Finally, as will be seen, much the same effect can be obtained from the following formulation, when player A takes action i for some small $\lambda > 0$, the ith propensity alone is updated thus

$$\begin{array}{ll} q^A_{in+1} = q^A_{in} + a_{ij} + \lambda\phi'_i(x) & \\ q^A_{kn+1} = q^A_{kn} & \text{for all } k \neq i. \end{array} \tag{16}$$

Here again $\phi(x)$ is a perturbation function satisfying the properties outlined in the context of stochastic fictitious play in Section 2. Under this rule, and given property (ii) of $\phi(x)$, if the probability of taking an action is low, then that action is strongly reinforced when taken.

The effect of the different specifications is easiest to see if we look at the effect on the rate of change of x (the underlying calculations are to be found in the proof of Proposition 2 below):

$$\begin{array}{rcl} E[x_{n+1}|q_n] - x_n & = & \frac{1}{Q^A_n}\Big(R(x_n)Ay_n + \lambda g^A(x_n, y_n)\Big) + O(\frac{1}{(Q^A_n)^2}) \\ E[y_{n+1}|q_n] - y_n & = & \frac{1}{Q^B_n}\Big(R(y_n)Bx_n + \lambda g^B(x_n, y_n)\Big) + O(\frac{1}{(Q^B_n)^2}). \end{array} \tag{17}$$

The expected motion of the stochastic process is still close to the replicator dynamics, but each equation now has an additional noise term depending on λ.

Given (14), it can be calculated that,

$$g^A_i(x_n, y_n) = \frac{x_n \cdot Ay_n - Nx_{in}(Ay_n)_i}{N-1}\,,\; g^B_j(x_n, y_n) = \frac{y_n \cdot Bx_n - My_{jn}(Bx_n)_j}{M-1} \tag{N1}$$

Thus, for example, if $x_{in} = 0$, then the expected change in x_{in} will be $\lambda x_n \cdot Ay_n/(N-1) > 0$. That is, the noise directs the system inward away from the boundary of $S_N \times S_M$. This has the consequence that every element of x will remain strictly positive. As for (15), it gives rise to,

$$g_i^A(x_n) = 1 - Nx_{in},\ g_j^B(y_n) = 1 - My_{jn} \tag{$N2$}$$

This specification of noise has been used in Gale et al. (1995) in the same manner as (17). Finally, (16) leads to

$$g^A(x_n) = R(x_n)\phi'(x_n),\ g^B(y_n) = R(y_n)\phi'(y_n) \tag{$N3$}$$

where $R(\cdot)$ is the replicator operator introduced in Section 3.

5 Stochastic Approximation

The principal means of analysis of stochastic models of learning in the recent literature has been by exploring the behaviour of associated deterministic systems, most often by employing the theory of stochastic approximation. The standard exposition, for example, Benveniste et al. (1990), of this theory assumes a discrete time stochastic process of the form

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) + \gamma_n^2 \epsilon(\theta_{n-1}, X_n). \tag{18}$$

The evolution of the parameter vector θ is determined by H and ϵ, an error term. X_n is a sequence of random vectors. The step size of the stochastic process is determined by $(\gamma_n)_{n\geq 0}$ a sequence of "small" scalar gains, with $\gamma_n \geq 0$, and $\sum_n \gamma_n = \infty$. If

$$\sum_n \gamma_n^\alpha < \infty \text{ for some } \alpha > 1$$

then we describe the algorithm as having *decreasing gain*. One obvious example of this is where $\gamma_n = 1/n$, and this is the most commonly analysed case. The following is referred to the mean or averaged ordinary differential equation (ODE) associated with (18),

$$\dot{\theta} = h(\theta), \tag{19}$$

where

$$h(\theta) = \lim_{n\to\infty} E[H(\theta, X_n)]$$

This is important in that recent results in the theory of stochastic approximation have shown the behaviour of this ODE (19) and of the stochastic process (18) are very closely linked. And indeed in this paper, the results obtained on the learning process will largely be obtained by analysis of the appropriate averaged ODE.

Given the expected motion of the reinforcement learning process (17), one might guess that the associated ODE would be the perturbed replicator dynamics

$$\dot{x} = R(x)Ay + \lambda g^A(x,y),\ \dot{y} = R(y)Bx + \lambda g^B(x,y). \tag{20}$$

Equally one might hope that in the absence of experimentation, the basic reinforcement learning model could be analysed using the replicator dynamics (8). Unfortunately, the situation considered here is rather more complex as the Erev-Roth model of reinforcement learning differs from the paradigm (18) above in two important respects. Firstly, the step size is endogenous, being determined by the accumulation of payoffs. Second, it is not scalar. There are two step sizes, $1/Q_n^A$ and $1/Q_n^B$, one for each player. It turns out that another variable, here denoted μ, is needed to take into account the relative speed of the learning of the two players.[7] What the following result establishes is that changes in relative speed do not change the steady states of the model, nor the behaviour of the model in the neighbourhood of those equilibria.

Proposition 2 *The ODE's associated with the reinforcement learning process generated by choice rule (1) and updating rules (14), (15) or (16) will be the following system*

$$\dot{x} = R(x)Ay + \lambda g^A(x,y),\ \dot{y} = \mu\left(R(y)Bx + \lambda g^B(x,y)\right),\ \dot{\mu} = \mu(x\cdot Ay - \mu y\cdot Bx). \tag{21}$$

If $(\hat{x},\hat{y})$ is an equilibrium for (20), then $(\hat{x},\hat{y},\hat{\mu})$, with $\hat{\mu} = \hat{x}\cdot A\hat{y}/\hat{y}\cdot B\hat{x}$ is an equilibrium for (21). Such an equilibrium is (un)stable for (20) if and only if it is (un)stable for (21).

Proof: In the Appendix. ∎

These are results on the local stability of deterministic ODE's. However, they are relevant to the stochastic learning processes we consider because of two types of results. Suppose that the linearisation of the ODE at an equilibrium point has at least one positive eigenvalue and so the equilibrium is unstable, the stochastic process converges to that point with probability zero (for details of such a result see, for example, Benaïm and Hirsch, 1999, Theorem 5.1). Second, suppose that the ODE has one or more asymptotically stable fixed points, then there is a positive probability of convergence of the stochastic learning process to any of these equilibria (see Benaïm, 1999, Theorem 7.3 for a result of this type).

In the case of stochastic fictitious play, it is rather easier to examine the discrete time stochastic process (10) in terms of an associated ODE. This is because for stochastic fictitious play the step size is exactly $1/n$ for both players. The associated ODE's are

$$\dot{u} = \overline{BR}(v) - u,\ \dot{v} = \overline{BR}(u) - v, \tag{22}$$

[7] Laslier et al. (2000) independently adopt a different method to overcome these problems. However, they only consider unperturbed reinforcement learning.

which I will refer to as the perturbed best response dynamics. These dynamics are also analysed in Benaïm and Hirsch (1999), Ellison and Fudenberg (2000), Hopkins (1999b), Hofbauer (2000) and Hofbauer and Hopkins (2000).

However, given Proposition 1, it is also clearly possible to find on the interior of $S_N \times S_M$ an ODE associated with the stochastic fictitious play process (13) which can be written

$$\dot{x} = \beta P(x)(Ay + \lambda\phi'(x)),\ \dot{y} = \beta P(y)(Bx + \lambda\phi'(y)). \tag{23}$$

As was noted in Section 3, this equation can be thought of as a noisy form of the replicator dynamics, multiplied by the factor β which increases the speed of learning. It is also possible to show that despite the apparent difference in form of these noisy replicator dynamics from the perturbed best response dynamics, they have many of the same properties.[8]

Proposition 3 *The two systems of ODE's, the perturbed best response dynamics, (22) and the noisy replicator dynamics (23), given the same perturbation function $\phi(\cdot)$, have the same fixed points, which share the same stability properties.*

Proof: In the Appendix. ∎

That is, many results obtained in terms of the process in historic frequencies can be carried over to the process in current choice frequencies. Furthermore, a link can be forged between stochastic fictitious play and reinforcement learning. For the exponential choice rule (6) where $\phi(x) = -\sum x_i \log x_i$, then in fact one can calculate that in fact $P(\cdot)$ is identical to the replicator operator $R(\cdot)$ so that the appropriate ODE's will be

$$\dot{x} = \beta R(x)\left(Ay - \lambda \log x\right),\ \dot{y} = \beta R(y)\left(Bx - \lambda \log x\right). \tag{24}$$

This together with Propositions 1 and 3 opens the intriguing prospect of analysing the behaviour of exponential fictitious play using the many existing results on the replicator dynamics. Clearly also, there is a similarity between (24) and the ODE's arising from perturbed reinforcement learning. This relationship will be explored in the next section.

6 Existence and Stability of Perturbed Equilibria

In this section, the qualitative behaviour of the two models of learning are compared, together with some evidence from experiments. First, it is established that

[8] Gaunersdorfer and Hofbauer (1995) obtain results in the opposite direction, showing a link between a best response dynamic in historical frequencies similar to (22) and the time averages of the replicator dynamics.

the equilibria of the perturbed forms of fictitious play and reinforcement learning can be identical. Second, these perturbed equilibria are close to but not identical to Nash equilibria. Furthermore, this difference helps to explain some experimental data, where there is no apparent convergence to Nash equilibrium even when unique.

For illustrative purposes, it will be useful to examine the class of 2×2 games that have a unique mixed strategy equilibrium. These games have attracted particular interest (Roth and Erev, 1998; Posch, 1997; Benaïm and Hirsch, 1999; Fudenberg and Kreps, 1993). Because we can replace x_2 by $1 - x_1$ for player A, and similarly for B, the state of the system can be summarised by the vector (x_1, y_1). Or in other words the learning dynamics will take place on the unit square. Without loss of generality, we can write the payoff matrices for the two players as

$$A = \begin{pmatrix} 1-a+c & c \\ c & a+c \end{pmatrix}, \; B = \begin{pmatrix} c & b+c \\ 1-b+c & c \end{pmatrix}, \tag{25}$$

where $1 > a, b > 0$ and $c > 0$. The latter constant ensures all payoffs are strictly positive. There is a unique mixed strategy equilibrium where $(x^*, y^*) = (b, a)$. However, the corresponding perturbed equilibrium of the perturbed replicator dynamics, either (20) or (21), is not in general at (x^*, y^*).

Note that for all the different specifications of noise introduced in Section 4, when $a = b = \frac{1}{2}$, that is when the mixed equilibrium is exactly in the middle of the unit square, the perturbed equilibrium $(\hat{x}, \hat{y})$ equals the Nash equilibrium (x^*, y^*). Otherwise the perturbed equilibrium may be some distance from the actual Nash equilibrium. Take one game investigated by Ochs (1995), which provides one of the data sets analysed by Erev and Roth. In this game, the Nash equilibrium was $\theta^* = (x_1^*, y_1^*) = (0.5, 0.1)$. This is illustrated in Figure 1. The arrows represent expected motion of learning and are generated under the simple assumption that a strategy whose expected return exceeds the other will grow in frequency. This is a property of many learning models, including the basic reinforcement model considered here, that is, without experimentation.

However, with experimentation the equilibrium is no longer θ^* but a perturbed equilibrium. In their first paper Roth and Erev (1995) a value of λ of 0.05 was used. In Erev and Roth (1998), the best fit of the data is obtained with a value of λ equal to 0.2. Given (N1) and a value of $\lambda = 0.05$ then the fixed point of the ODE (20), solving these cubic equations numerically, will be $\hat{\theta}_1 = (0.6405, 0.1063)$. With a value of $\lambda = 0.2$ then $\hat{\theta}$ moves to $\hat{\theta}_2 = (0.7428, 0.2673)$, though, as λ increases and the noise dominates, the equilibrium will move toward (0.5,0.5). Points $\{1, 2, 3, 4\}$ represent aggregate data in blocks of 16 periods from the experiments run by Ochs (1995) as reported in McKelvey and Palfrey (1995).

Hence, if this model of reinforcement learning accurately describes subjects' learning behaviour then Nash equilibrium is not to be expected. Even if in the long run the learning process converges, it will be to the perturbed equilibrium $(\hat{x}, \hat{y})$. Note

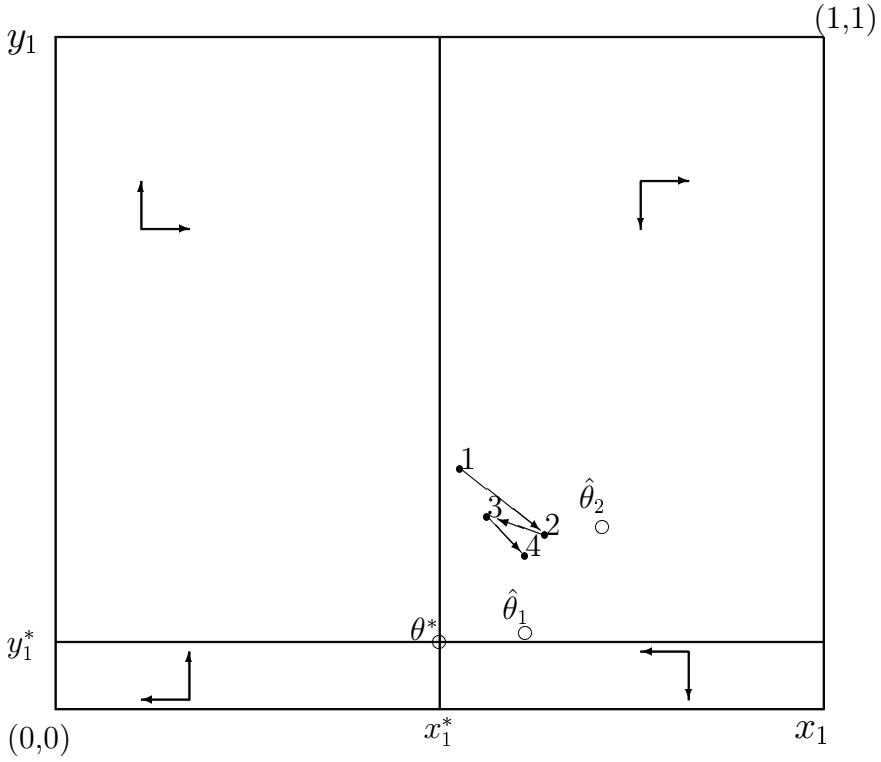

Figure 1: Perturbed equilibria and game dynamics.

that the motion "away" from Nash equilibrium that Roth and Erev find in their data is toward this perturbed equilibrium. However, this model is not unique in possessing an equilibrium which is not identical to Nash.

This is also a characteristic of the stochastic version of fictitious play. An equilibrium for these dynamics, like for the perturbed replicator dynamics, can be some distance away from the Nash equilibrium of the underlying game, the distance depending on the value of the noise parameter λ (see, for example, Fudenberg and Levine, 1998, pp108-9). The following result shows that the resemblance is not coincidental: a fixed point for perturbed reinforcement learning is a fixed point for stochastic fictitious play and vice versa.[9]

Proposition 4 *A point $(\hat{x}, \hat{y})$ is an equilibrium point for the perturbed best response dynamics (22), if and only if $(\hat{x}, \hat{y}, \hat{\mu})$ is an equilibrium point for the perturbed replicator dynamics (21) with noise (N3). Let $(\hat{x}, \hat{y}, \hat{\mu})$ be an equilibrium point for (21) with noise specification (N2). Then, there exists a suitable perturbation function $\phi(\cdot)$*

[9]Roth and Erev's actual specification of noise, because there g^A, g^B depend on both x and y, does not fit this pattern, though as we have seen the perturbed equilibria it produces are qualitatively similar.

such that $(\hat{x}, \hat{y})$ *is an equilibrium point for (22).*

Proof: In the Appendix. ∎

Equally McKelvey and Palfrey (1995) propose a new equilibrium concept, "quantal response equilibrium" or QRE, which is based on perturbation of players' payoffs. For the game considered here any QRE would like $(\hat{x}, \hat{y})$ be up and right of the Nash equilibrium. McKelvey and Palfrey in fact estimate the QRE from Ochs' experimental data at (0.649,0.254) which is roughly intermediate between $\hat{\theta}_1$ and $\hat{\theta}_2$. Indeed, any QRE is an equilibrium point of the stochastic fictitious play process (22) above. At the basis of the concept of QRE (McKelvey and Palfrey, 1995) is the construction of perturbed best response functions in the same manner as they are in stochastic fictitious play, that is, as the solution of a perturbed maximisation problem. Indeed the most commonly used form of perturbed best response function is the same in both cases: the exponential or logistic form (6). A QRE is a point at which every player plays a perturbed best response to the play of other players. Such a point is clearly a fixed point also of the dynamics (22) given the same perturbed best response function is used in both cases.[10]

Returning to the relationship between reinforcement learning and stochastic fictitious play, we have seen that the two models share rest points. But it is also possible to establish that these equilibria have similar stability properties under the two different learning processes. The statement of the result below is different for stability and instability. This is because for λ large enough, any mixed equilibrium will be stable under perturbed dynamics as the noise swamps all other factors. Second, there is no result here for pure equilibria. As noted perturbed equilibria are not identical to Nash. In fact, Nash equilibria on the boundary of the simplex may have no corresponding perturbed equilibrium at all (see for example, Binmore and Samuelson, 1999). Consequently, it is difficult to establish any general result stronger than perturbed equilibria corresponding to strict Nash equilibria are asymptotically stable for small enough λ.

Proposition 5 *If a perturbed mixed equilibrium is asymptotically stable for all perturbed best response dynamics (22), then it is asymptotically stable for the noisy replicator dynamics (21) specification (N2) or (N3). If, for* λ *sufficiently small, a perturbed mixed equilibrium is unstable for all perturbed best response dynamics (22), then there is a* $\lambda > 0$ *such that it is unstable for the noisy replicator dynamics (21) specification (N2) or (N3).*

Proof: In the Appendix. ∎

[10] Unfortunately, the literature on QRE and that on learning have up to this point had little overlap.

In the special case of 2×2 games with a unique mixed equilibrium, it is possible to go further and consider global convergence. This is important in that Erev and Roth (1998) use data from experimental play of this class of games to test between reinforcement learning and stochastic fictitious play. However, it is possible to show that a noisy reinforcement learning model converges to a perturbed equilibrium, just as Fudenberg and Kreps (1993), and Benaïm and Hirsch (1999) have shown for stochastic fictitious play. That is, the two rival models have the same long run properties. It is first useful to establish a preliminary result.

Proposition 6 *The perturbed equilibrium $(\hat{x}, \hat{y})$ of the game (25) is globally asymptotically stable under the perturbed replicator dynamics (20), for noise specification (N1), (N2) and (N3).*

Proof: If the divergence, defined as $\sum_i \frac{\partial \dot{x}_i}{\partial x_i} + \sum_j \frac{\partial \dot{y}_j}{\partial y_j}$, is negative, this is sufficient in a 2 dimensional system to rule out cycles or other exotic behaviour. The replicator dynamics (8) have zero divergence on $S_2 \times S_2$, (Hofbauer and Sigmund, 1998, pp132-3). To see this, consider the modification of (8) $\dot{x}/V$, $\dot{y}/V$ where $V(x, y) = x_1 x_2 y_1 y_2$. Note that a positive transformation, such as division by V, only changes the velocity not the orbits of a dynamical system. Thus, as the modified system has zero divergence, so does (8). It can be calculated that for noise of form (N1), (N2) and (N3) in the 2×2 case, $\frac{\partial (g_1^A / V)}{\partial x_1} < 0$ and $\frac{\partial (g_1^B / V)}{\partial y_1} < 0$. This implies that the divergence is negative. Hence, as the perturbed replicator dynamics (20) are simply (8) plus a term with negative divergence, (20) will have negative divergence. Thus, the flow of the perturbed dynamics (20) must be volume contracting on the whole of $S_2 \times S_2$ and converge to the unique equilibrium point. ∎

Note that unfortunately this result does not imply that perturbed equilibrium is a global attractor for the more complex system (21). More specifically, because of the extra equation in (21), the system is no longer two-dimensional and negative divergence is no longer sufficient for global convergence. Global stability results are in general very difficult to obtain for systems of greater than two dimensions.[11] Hence, the only way to obtain an analytic result is to modify the learning model to make it more tractable. This is the methodology adopted in Arthur (1993) and Posch (1997).

Assumption A: Normalisation. At each period, for each player k after propensities are updated by the addition of payoffs according to any reinforcement updating rule, every propensity is multiplied by an appropriate factor so that $Q_n^k = Q_0 + n$ for $k = A, B$, but leaving x_n, y_n unchanged.

The normalised process has a unique deterministic step size of order $1/n$. This

[11] Hofbauer and Hopkins (2000), however, give some results on global convergence for stochastic fictitious play for games larger than 2×2. Duffy and Hopkins (2000) show global convergence of reinforcement learning in one class of games.

allows easy application of existing results in the theory of stochastic approximation. For example, one can show the following.

Proposition 7 *The ODE's associated with the perturbed reinforcement learning process with noise specification (N1) or (N2) or (N3), and under Assumption A, will be the noisy replicator dynamics (20). Hence, in game (25), all of these stochastic learning processes converge with probability one to a perturbed equilibrium* $(\hat{x}, \hat{y})$.

Proof: In the Appendix. ∎

This result shows that while normalisation permits additional formal results it leaves much unchanged. Proposition 7 together with the earlier Proposition 2 establishes that the imposition of Assumption A does not change the rest points of the learning process or their local stability properties. It may make some difference in learning behaviour away from equilibrium but it is difficult to assess the significance of this.[12] One other consequence of the imposition of Assumption A is that the expected motion of reinforcement learning can be expressed as a simple multiple of that of stochastic fictitious play.

Corollary 8 *The ODE's (20) associated with the perturbed reinforcement learning process with noise specification (N3) and* $\phi(x) = -\sum x_i \log x_i$ *under Assumption A are identical to a positive factor with the ODE's (24) associated with exponential fictitious play.*

This follows from inspection of (24), (20) and (N3). The intuition is simply that as the ODE associated with stochastic fictitious play can be written as a form of perturbed replicator dynamic, to match reinforcement learning with stochastic fictitious play, one merely needs to find the right noise function.

7 Conclusions

This paper investigates models of reinforcement learning that discard useful information and which employ no optimisation. It is shown that they can generate exactly the same asymptotic results as stochastic fictitious play. The link between the two models is even stronger if attention is confined to local stability. However, it remains clear that the two models are not identical. Both Erev and Roth (1998) and Camerer and Ho (1999) have used experimental data to test between stochastic fictitious play

[12]For example, simulation suggests that even without normalisation the perturbed equilibrium in games of type (25) is still a global attractor for perturbed reinforcement learning, but it has not been possible to establish a proof.

and reinforcement learning. What the analysis presented here suggests is that the only way that learning behaviour generated by the two models may differ is in speed of passage along similar paths. This difference in speed may, however, be significant. Indeed, depending on the the degree of optimisation present, here captured by the parameter β, stochastic fictitious play can be arbitrarily faster than reinforcement learning. These differences may well be important relative to the timescale of experiments and therefore be identifiable econometrically.

Reinforcement learning and stochastic fictitious play differ along two axes, use of information and optimisation. The role of optimisation is clearly identifiable: from Proposition 8 the expected motion of stochastic fictitious play is exactly β times faster than that of reinforcement learning. However, that leaves the puzzle that there is no apparent equivalent role for hypothetical reinforcement. Cheung and Friedman (1997) fit a stochastic fictitious play model to data from experiments where the information given to participants varied. Their estimates of β, which indicates the level of optimisation, were higher in the full information treatments, perhaps reflecting the subjects' greater confidence in these environments. So, there is evidence that in practice empirical estimates of the parameter β capture both factors.

The results presented here are in the context of strategic form games. So one might hypothesise that in games with a non-trivial extensive form and hence where the manipulation of information and the forming of hypotheticals are more important, the two models of learning might lead to quite different outcomes. This is the rationale for the experiments reported in Feltovich (2000) on a constant sum game with asymmetric information. However, he reports that even here the "two models yield qualitatively similar patterns of behavior", which is, of course, very much in concordance with the results in this paper. This seems to imply somewhat surprisingly that the possession of or the lack of information about opponents' actions, and the degree of players' sophistication, may have no effect on the asymptotic outcome of learning.

However, this paper does not claim that any and all information that agents might receive is irrelevant. Rather, it is possible that the debate over fictitious play and reinforcement learning about whether or not agents use hypothetical reinforcement has been a debate about the wrong type of information. For example, there is some experimental evidence that giving information about an opponent's payoffs, as opposed to her moves, can have quite a substantial impact. This is because it allows players to identify what is a dominated strategy for their opponents (see for example, Cooper, Garvin and Kagel, 1997; Dekel, Fudenberg, and Levine, 1999). However, this is a form of game theoretic reasoning not captured by either model considered here.[13] Second, the two models considered here are concerned with pairs of agents

[13]However, clearly belief-based models such as fictitious play are more easily modifiable to take such reasoning into account. One can impose the refinement that agents' prior beliefs must place zero weight on strategies that are dominated for their opponents. Again, see Cooper et al. (1997) and Dekel et al. (1999).

playing in isolation. In contrast, there have been experiments, for example, Huck et al. (1999) and Duffy and Feltovich (1999) where some subjects are informed about the behaviour of other subjects in a way that permits learning by imitation. Here, it seems that whether this information is provided or withheld can have significant effects on play.

Appendix

Calculations from Section 3.1: Now, if in period n event ij occurs, then the change in x_i will be

$$x_{in+1} - x_{in} = \frac{q_{in}^A + a_{ij}}{Q_n^A + a_{ij}} - \frac{q_{in}^A}{Q_n^A} = \frac{(1 - x_{in})a_{ij}}{Q_n^A + a_{ij}} = \frac{(1 - x_{in})a_{ij}}{Q_n^A} + O(\frac{1}{(Q_n^A)^2}).$$

But if the event kj occurs for $k \neq i$, the change in x_i will be

$$x_{in+1} - x_{in} = \frac{q_{in}^A}{Q_n^A + a_{kj}} - \frac{q_{in}^A}{Q_n^A} = \frac{-x_{in}a_{kj}}{Q_n^A + a_{kj}} = \frac{-x_{in}a_{kj}}{Q_n^A} + O(\frac{1}{(Q_n^A)^2}).$$

Given that event ij occurs with probability $x_{in}y_{jn}$ and event kj with probability $x_{kn}y_{jn}$, one can calculate

$$E[x_{in+1}|q_n] - x_{in} = \frac{1}{Q_n^A}(-x_{1n}x_{in}, ..., x_{in}(1 - x_{in}), ..., -x_{Nn}x_{in}) \cdot Ay_n + O(\frac{1}{(Q_n^A)^2}).$$

But given the definition of the replicator operator $R(\cdot)$, we have arrived at (9).

Proof of Proposition 1. Starting with (12), the first step is to differentiate the first order conditions (5). From this one can obtain $d\overline{BR}(v)/dAv = -(\phi'')^{-1}(\cdot)A/\lambda$ which can be written more compactly as $\beta P(x)A$. As Hopkins (1999b) notes, $P(x)$ is a symmetric matrix, which is positive definite with respect to $\mathbb{R}_0^N$, where $\mathbb{R}_0^N = \{x \in \mathbb{R}^N : \sum x = 0\}$. That is, $z \cdot P(x)z > 0$ for all $z \in \mathbb{R}_0^N$. If one in addition notes that the expected change in v_n is equal to $(\overline{BR}(u_n) - v_n)/(n+1) = (y_n - v_n)/(n+1)$, it is possible to see that in vector form,

$$E[x_{n+1}|x_n, y_n] - x_n = \frac{\beta}{n+1}P(x_n)\,(Ay_n - Av_n) + O(\frac{1}{n^2}).$$

The final step is to write the expected motion entirely in terms of current choice probabilities. Note that from the first order conditions (5), one can substitute $-Av_n = \lambda\phi'(x_n)$. ∎

Proof of Proposition 2. First, I show the calculations which generate the expected motion of the reinforcement learning process given the updating rule (15). They are easily extended to the other rules (14) and (16). If in period n event ij occurs, then the change in x_i will be

$$x_{in+1} - x_{in} = \frac{q_{in}^A + a_{ij} + \lambda}{Q_n^A + a_{ij} + N\lambda} - \frac{q_{in}^A}{Q_n^A} = \frac{(1 - x_{in})a_{ij} + \lambda(1 - Nx_{in})}{Q_n^A + a_{ij} + N\lambda}.$$

But if the event kj occurs for $k \neq i$, the change in x_i will be

$$x_{in+1} - x_{in} = \frac{q_{in}^A + \lambda}{Q_n^A + a_{kj} + N\lambda} - \frac{q_{in}^A}{Q_n^A} = \frac{-x_{in}a_{kj} + \lambda(1 - Nx_{in})}{Q_n^A + a_{kj} + N\lambda}.$$

Given that event ij occurs with probability $x_{in}y_{jn}$ and event kj with probability $x_{kn}y_{jn}$, one can calculate

$$E[x_{in+1}|q_n] - x_{in} = \frac{1}{Q_n^A}\left(R(x_n)Ay_n + \lambda(1 - Nx_{in})\right) + O(\frac{1}{(Q_n^A)^2}).$$

Next, the step size for the overall stochastic learning process is set to the step size of the first player, that is, $\gamma_n = 1/Q_n^A$. Note that if all payoffs are bounded and strictly positive, with probability one, $\lim_{n\to\infty}\gamma_n = 0$, $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$. Then, a new variable is introduced to take into account the relative speed of player B's learning. Define $\mu_n = Q_n^A/Q_n^B$. Again if all payoffs are bounded and strictly positive, μ_n is also bounded and strictly positive even as n goes to infinity. We can write the stochastic process as

$$\begin{aligned}
x_{n+1} - x_n &= \frac{1}{Q_n^A}\left(R(x_n)Ay_n + \lambda g^A(x_n, y_n) + \eta^A(X_n)\right) + O(\frac{1}{(Q_n^A)^2}), \\
y_{n+1} - y_n &= \frac{1}{Q_n^A}\mu_n\left(R(y_n)Bx_n + \lambda g^B(x_n, y_n) + \eta^B(X_n)\right) + O(\frac{1}{(Q_n^B)^2}), \\
\mu_{n+1} - \mu_n &= \frac{1}{Q_n^A}\mu_n\left(x_n \cdot Ay_n - \mu_n y_n \cdot Bx_n + \eta^\mu(X_n)\right) + O(\frac{1}{(Q_n^B)^2}). \qquad (26)
\end{aligned}$$

The variables η^k for $k = A, B, \mu$, each have expectation zero. Note that here X_n is simply the indicator function giving the outcome (out of the $N \times M$ possible) of the two players' randomisations in period n. To construct the ODE's (19) associated with this system, one takes the average over the possible realisations to obtain (21).

The resultant dynamics (21) are on $S_N \times S_M \times \mathbb{R}_+$. If $(\hat{x}, \hat{y})$ is an equilibrium for (20), I hope it is clear that $(\hat{x}, \hat{y}, \hat{x} \cdot A\hat{y}/\hat{y} \cdot B\hat{x})$ is an equilibrium for (21).[14] Taking the linearisation at such a perturbed equilibrium point $(\hat{x}, \hat{y}, \hat{\mu})$, one obtains

$$K = \left(\begin{array}{ccc} J & & 0 \\ \hline \mu(Ay - \mu yB) & \mu(xA - \mu Bx) & -x \cdot Ay \end{array}\right),$$

where J is the Jacobian matrix of the simpler system (20). Note that, writing $z = (z_1, z_2)$ with $z_1 \in \mathbb{R}^{N+M}$ and z_2 scalar, the eigenvalue equation for the above matrix, that is, $Kz = \chi z$, for some eigenvalue χ, can be decomposed into two separate equations, $Jz_1 = \chi z_1$, and $(\mu(Ay - \mu yB), \mu(xA - \mu Bx)) \cdot z_1 - x \cdot Ayz_2 = \chi z_2$. Hence

[14] It is true that there is another equilibrium at $(\hat{x}, \hat{y}, 0)$. But it is easy to establish using the arguments in this proof that such an equilibrium will always be unstable.

$N + M$ of the eigenvalues of K are the eigenvalues of the matrix J. The remaining eigenvalue is therefore $-x \cdot Ay$. In conclusion, at an equilibrium $(\hat{x}, \hat{y}, \hat{\mu})$ of the system (21) there is an additional negative eigenvalue relative to the linearisation J at an equilibrium $(\hat{x}, \hat{y})$. If J has any positive eigenvalues, K has too, and the perturbed equilibrium is unstable. If J has all negative eigenvalues, so does K. ∎

Proof of Proposition 3. A fixed point of the dynamics (22) in historical frequencies is where the first order conditions (5) are simultaneously satisfied for both players, that is,

$$Ay + \lambda\phi'(x) = 0,\ Bx + \lambda\phi'(y) = 0. \tag{27}$$

It is clear that every point that satisfies (27) is a fixed point for the ODE's in choice probabilities (23). Furthermore, given the positive definiteness of $P(x)$ and $P(y)$, established in Proposition 1, these are the only fixed points of the ODE's.

Turning now to stability properties, the first step is to construct the linearisation of the dynamics. In the case of the noisy replicator dynamics (23) associated with the stochastic fictitious play process in current choice probabilities, differentiate at a perturbed equilibrium point $\hat{\theta} = (\hat{x}, \hat{y})$ to obtain

$$\frac{d\dot{x}}{dx} = \beta\left(P'(Ay + \lambda\phi'(x)) + P\lambda\phi''\right) = -I$$

given that $P(x) = -(\phi'')^{-1}$. Equally,

$$\frac{d\dot{x}}{dy} = \beta P(x)A.$$

Thus the linearisation can be written

$$\beta\begin{pmatrix} P(x) & 0 \\ 0 & P(y) \end{pmatrix}\begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} - I = \beta P(\hat{\theta})C - I. \tag{28}$$

Now, to construct the equivalent linearisation for the dynamics in historical frequencies, it is enough to look at the proof of Proposition 1 to see that, for example, $d\overline{BR}(v)/dv$ can be written as $\beta P(x)A$ and hence the linearisations are identical. Looking at the linearisation (28), it is clear that its eigenvalues will be of the form $\beta\chi - 1$ where χ is some eigenvalue of the matrix PC. Hence, for generic values of the parameter β, all perturbed equilibria will be hyperbolic. Hence, their local stability properties will be entirely determined by the appropriate linearisation. ∎

Proof of Proposition 4. From proposition 2, it is possible to work with the simpler system (20) as it has the same equilibria as the ODE's (21) associated with perturbed reinforcement learning. Given (N3), the perturbed replicator dynamics (20) can be rewritten as

$$\dot{x} = R(x)\left(Ay + \lambda\phi'(x)\right),\ \dot{y} = R(y)\left(Bx + \lambda\phi'(y)\right). \tag{29}$$

These equations clearly have fixed points that satisfy the conditions (27) and given the positive definiteness of $R(\cdot)$ these are the only class of fixed points. But these

are exactly the first order conditions that define a fixed point for the perturbed best response dynamics. So for an identical perturbation function $\phi(x)$, a point $(\hat{x}, \hat{y})$ is an equilibrium for (22) if and only if it is an equilibrium for (20).

In the case of noise (N2), note that the perturbed replicator dynamics) can be written

$$\dot{x} = R(x)Ay + \lambda(u - Nx),\ \dot{y} = R(y)Bx + \lambda(u - My). \tag{30}$$

where u is a vector of ones, and $R(x)$ and $R(y)$ are replicator operators. The functions, $\phi(x) = \sum_{i=1}^{N} \log x_i$, $\phi(y) = \sum_{i=1}^{M} \log y_i$ satisfy the 2 conditions set out in Section 2 to act as suitable perturbations to construct a perturbed best response function. Second note that

$$\phi'(x) \cdot R(x) = u - Nx.$$

Combining this with (30), we again obtain (29), but this time with $\phi(x) = \sum \log x_i$ and $\phi'(x) = (1/x_1, ..., 1/x_N)$. Given that $R(x)$ and $R(y)$ are positive definite, the fixed points of the dynamic again must satisfy the simultaneous equations (27) above. ∎

Proof of Proposition 5. First, from Proposition 2, the equilibria and their stability properties under the two systems (20) and (21) are the same. So in this proof, I work with the simpler system (20).

It was shown in the proof of Proposition 4 that the perturbed replicator dynamics for both (N2) and (N3) could be written in the form (29). At a perturbed equilibrium $\hat{\theta} = (\hat{x}, \hat{y})$ satisfying the conditions (27), the linearisation of these dynamics can be written

$$J = R(\hat{\theta})C + \lambda R(\hat{\theta})\Phi(\hat{\theta}) = \begin{pmatrix} 0 & R(\hat{x})A \\ R(\hat{y})B & 0 \end{pmatrix} + \lambda \begin{pmatrix} R(\hat{x})\phi''(\hat{x}) & 0 \\ 0 & R(\hat{y})\phi''(\hat{y}) \end{pmatrix}. \tag{31}$$

Now when $\phi(x) = -\sum x_i \log x_i$ as it does for the exponential version of fictitious play (6), then first, $P(\cdot) = R(\cdot)$, and, second, $R(x)\phi''(x) = -I$ and similarly $R(y)\phi''(y) = -I$. Hence (31) gives the same linearisation (to the positive factor β) as in (28) above.

For more general perturbation functions $\phi(\cdot)$, the aim is to show that the eigenvalues of J, the linearisation of the perturbed replicator dynamics have the same sign pattern at any perturbed equilibrium as the eigenvalues of the linearisation of the perturbed best response dynamics, given in (28) above. If $P(\hat{\theta})C$ has at least one positive eigenvalue for all suitable P, then the perturbed equilibrium $\hat{\theta}$ will be unstable under all perturbed best response dynamics for sufficiently small λ. Hence $R(\hat{\theta})C$ has at least one positive eigenvalue and so does J for small enough λ. Because PC has a zero trace, it has either both positive and negative eigenvalues or all with zero real part. Hence, an equilibrium can only be asymptotically stable for all perturbed best response dynamics if $P(\hat{\theta})C$ has all eigenvalues with real part zero. Now, PC has all eigenvalues with zero real part for for all suitable P, and nonzero C, if and only if (A, B) is a rescaled zero sum game (Hofbauer and Hopkins, 2000). Then

$\xi \cdot A\eta + c\eta \cdot B\xi = 0$ for some $c > 0$ and for any $\xi \in I\!R_0^N$ and $\eta \in I\!R_0^M$ (Hofbauer and Sigmund, 1998, p128-9). Note that if we multiply B and $\phi''(y)$ by the appropriate positive constant, c, and divide $R(y)$ by c, J is unchanged. However, now after this rescaling $C + C^T = 0$. Note that as $\phi''(\cdot)$ is negative definite by property (i) in Section 2, so is Φ. Hence, $C + \lambda\Phi$ is negative definite and consequently (see eg Hopkins, 1999a, Lemma 2) $R(C + \lambda\Phi)$ has all eigenvalues with real part negative. ∎

Proof of Proposition 7. Let $\pi_n^k = q_{n+1}^k - q_n^k$, that is, the increment to a player's propensities. For example, under updating rule (15), given the event ij, then $\pi_{in}^A = a_{ij} + \lambda$. Then, normalisation involves that after updating every propensity of player k is multiplied by a factor $(Q_0 + n + 1)/(Q_n^k + \sum_i \pi_{in}^k)$. One can check that (x_n, y_n) are unchanged under this transformation, but that each Q_{n+1}^k is renormalised to $Q_0 + n + 1$. There is therefore a unique step size equal to $1/(Q_0 + n)$. The additional factor μ is now constant and equal to one. Thus we can discard the third equation from (26). Take what is left, average over all possible events ij and one now obtains (20). The global stability of $(\hat{x}, \hat{y})$ under (20) was established in Proposition 6. This is sufficient to prove convergence with probability one of the associated stochastic process given a step size of $O(1/n)$. See for example, Benveniste et al. (1990, p46, Corollary 6) or Benaïm and Hirsch (1999, Theorem 3.3). ∎

References

Arthur, W.B. (1993). "On designing economic agents that behave like human agents," *J. Evol. Econ.*, **3**, 1-22.

Benaïm, M. (1999). "Dynamics of stochastic algorithms," in *Séminaire de Probabilités XXXIII*, J. Azéma et al. Eds, Berlin: Springer-Verlag.

Benaïm, M., Hirsch, M.W. (1999). "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games," *Games Econ. Behav.*, **29**, 36-72.

Benveniste, A., Métivier, M., Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* Berlin: Springer-Verlag.

Binmore, K., Samuelson, L. (1999). "Evolutionary drift and equilibrium selection," *Rev. Econ. Studies*, **66**, 363-393.

Blume, A., DeJong, D.V., Neumann, G.R., Savin, N.E. (2000). "Learning and communication in sender-receiver games: an econometric investigation," working paper, University of Iowa.

Börgers, T., Sarin, R. (1997). "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, **77**, 1-14.

Camerer, C., Ho, T-H. (1999). "Experience-weighted attraction learning in normal form games," *Econometrica*, **67**, 827-874.

Cheung, Y-W., Friedman, D. (1997). "Individual learning in normal form games: some laboratory results," *Games Econ. Behav.*, **19**, 46-76.

Cooper, D.J., Garvin, S., Kagel, J.H. (1997). "Signalling and adaptive learning in an entry limit pricing game," *Rand J. of Econ.*, **28**, 662-83.

Dekel, E., Fudenberg, D., Levine, D.K. (1999). "Payoff information and self-confirming equilibrium," *J. Econ. Theory*, **89**, 165-85.

Duffy J., Feltovich, N. (1999). "Does observation of others affect learning in strategic environments? An experimental study," *Int. J. Game Theory*, **28**, 131-152.

Duffy J., Hopkins, E. (2000). "Learning, information and sorting in market entry games: theory and evidence," working paper, Universities of Pittsburgh and Edinburgh.

Ellison, G., Fudenberg, D. (2000). "Learning purified mixed equilibria," *J. Econ. Th.*, **90**, 84-115.

Erev, I., Roth, A.E. (1998). "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *Amer. Econ. Rev.*, **88**, 848-881.

Feltovich, N. (2000). "Reinforcement-base vs. belief-based learning models in experimental asymmetric-information games," *Econometrica*, **68**, 605-41.

Fudenberg, D., Kreps D. (1993). "Learning mixed equilibria," *Games Econ. Behav.*, **5**, 320-367.

Fudenberg, D., Levine D. (1998). *The Theory of Learning in Games.* Cambridge, MA: MIT Press.

Gale, J., Binmore, K., Samuelson, L. (1995). "Learning to be imperfect: the ultimatum game," *Games Econ. Behav.*, **8**, 56-90.

Gaunersdorfer, A., and Hofbauer, J. (1995). "Fictitious play, Shapley polygons, and the replicator equation," *Games Econ. Behav.*, **11**, 279-303.

Hofbauer, J. (2000). "From Nash and Brown to Maynard Smith: equilibria dynamics and ESS," forthcoming *Selection.*

Hofbauer, J., Hopkins, E. (2000). "Learning in perturbed asymmetric games," working paper.

Hofbauer, J., Sigmund, K. (1998). *Evolutionary Games and Population Dynamics.* Cambridge, UK: Cambridge University Press.

Hopkins, E. (1999a). "Learning, matching and aggregation," *Games Econ. Behav.*, **26**, 79-110.

Hopkins, E. (1999b). "A note on best response dynamics," *Games Econ. Behav.*, **29**, 138-150.

Huck S., Norman, H., Oechssler, J. (1999). "Learning in Cournot oligopoly–an experiment," *Econ. J.*, **109**, pp. C80-95.

Laslier, J-F., Topol, R., Walliser, B. (2000). "A behavioral learning process in games," forthcoming *Games Econ. Behav.*

McKelvey, R.D., Palfrey, T.R. (1995). "Quantal response equilibria for normal form games," *Games Econ. Behav.*, **10**, 6-38.

Ochs, J. (1995). "Simple games with unique mixed strategy equilibrium: an experimental study," *Games Econ. Behav.*, **10**, 202-217.

Posch, M. (1997). "Cycling in a stochastic learning algorithm for normal form games," *J. Evol. Econ.*, **7**, 193-207.

Roth, A.E., Erev, I. (1995). "Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term," *Games Econ. Behav.*, **8**, 164-212.

Rustichini, A. (1998). "Optimal properties of stimulus-response learning models," *Games Econ. Behav.*, **29**, 244-73.

Salmon, T. (1999). "An evaluation of econometric learning models of adaptive learning," working paper, California Institute of Technology.

Sarin, R., Vahid, F. (1998). "Predicting how people play games: a procedurally rational model of choice," Working Paper, Texas A&M University.

Van Huyck, J.B., Battalio, R.C., Rankin, F.W. (1997). "On the origin of convention: evidence from coordination games," *Econ. J.*, **107**, 576-596.

Vriend, N.J. (1997). "Will reasoning improve learning?" *Econ. Letters*, **55**, 9-18.